Exploring the Similarity between Social Knowledge Sources and Twitter for Cross-domain Topic Classification

Andrea Varga, Amparo Elizabeth Cano and Fabio Ciravegna[1]

OAK Group,
Dept. of Computer Science,
The University of Sheffield,
United Kingdom
firstinitial.lastname@dcs.shef.ac.uk

Abstract. The rapid rate of information propagation on social streams has proven to be an up-to-date channel of communication, which can reveal events happening in the world. However, identifying the topicality of a short messages (e.g. tweets) distributed on these streams poses new challenges in the development of accurate classification algorithms. In order to alleviate this problem we study for the first time a transfer learning setting aiming to make use of two frequently updated social knowledge source (KS) (DBpedia and Freebase) for detecting topics in tweets. In this paper we investigate the similarity (and dissimilarity) between these KS and Twitter at the lexical and conceptual(entity) level. We also evaluate the contribution of these types of features and propose various statistical measures for determining the topics which are highly similar or different in KS and tweets. Our findings can be of potential use to machine learning or domain adaptation algorithms aiming to use named entities for topic classification of tweets. These results can also be valuable in the identification of representative sets of annotated articles from the KS, which can help in building accurate topic classifiers for tweets.

Keywords: social knowledge sources, transfer learning, named entities, data analysis

1 Introduction

Micropost platforms such as Twitter serve as a real-time channel of information regarding events happening around the world. Compared to traditional news sources, microposts communicate more rapidly up-to-date information on a large number of topics. Identifying these topics in real-time could aid in different scenarios including i.e., emergency response, and terrorist attacks.

However, microposts mining poses several challenges since some of the characteristics of a tweet include: i) *the use of non-standard English*; ii) *the restricted size of a post (limited to 140 characters)*; iii) *the frequent misspellings and use of jargon*; and iv) *the frequent use of abbreviations*.

The dynamic changes in both vocabulary and style pose additional challenges for supervised classification algorithms, since the collection of annotated data becomes particularly difficult. However, frequently updated social knowledge sources(KS), such as DBpedia and Freebase, present an abundant source of structured data which could potentially aid in streamed topic detection. Similar to Twitter, these sources exhibit the following characteristics: i) *they are constantly edited by web users*; ii) *they are social and built on a collaborative manner*; iii) *they cover a large number of topics*; and iv) *they provide plentiful amount of annotated data*.

In this work we present *for the first time* a comparative study which analyses the similarity between Twitter and two frequently updated KS including *DBPedia* and *Freebase*. This comparative study includes the analysis of various cross-domain(CD) topic classifiers built on these KSs considering different lexical and conceptual features derived from named entities. Our intuition for the conceptual features is that the mention of certain entity types could be a good indicator for a specific topic. For e.g. a tweet containing the entity "Obama" is more likely to be a trigger for the topics "Politics" and "War&Conflict" than for the topic "Entertainment". Similarly, "Lady Gaga" is more likely to appear in tweet messages about the topics "Entertainment" or "Music", than about the topic "Sports".

In addition, we propose different statistical measures for quantifying the similarity and differences between these KS and tweet messages. The main research questions we investigate are the following: *i*) *Do KSs reflect the lexical changes in Twitter?*; *ii*) *Which features make the KSs look more similar to Twitter?*; *iii*) *How similar or dissimilar are KS to Twitter*; and *iv*) *Which similarity measure does better quantify the lexical changes between KS and Twitter?*

The main contributions of this paper are as follows: *i*) *we present a methodology for building CD topic classifiers for tweets making use of KSs*; and *ii*) *we present a comparative analysis exploring the similarity between KSs and Twitter at the level of words and named entities for CD topic classification*;

In the remaining of the paper we briefly describe the DBpedia and Freebase KS, we then present the state-of-the-art approaches in topic classification of Tweets, then we describe the main methodology and present the results obtained.

2 Social Knowledge Sources: an overview of DBpedia and Freebase

In this section we briefly review the main features of the DBpedia and Freebase KSs, highlighting the differences and similarities between them.

DBpedia[1] is a structured knowledge base derived from Wikipedia[2], the largest collaboratively maintained encyclopaedia. The latest released, DBpedia 3.7, classifies 1.83 million resources into 740,000 Wikipedia categories and 18,100,000 YAGO2 categories. For a given Wikipedia article DBpedia provides the following information [4]: *i*) the *title* of the Wikipedia article; *ii*) the *abstract* of the article corresponding to the first few

[1] http://dbpedia.org
[2] http://wikipedia.org

paragraphs containing up to 500 words; *iii*) the Wikipedia *categories* (topics) assigned to the article; *iv*) various links such as the *external links* pointing to external Web resources, *redirects* pointing to other articles about synonymous terms, *pagelinks* describing all the links in the article, *inter-language links* pointing to the translations of the article into multiple languages; *v*) *disambiguation pages* explaining different meaning of homonyms about a given term; *vi*) *images* depicting the resources from the article; *vii*) *homepage* or *website* information for an entity such as organisation or company; and *viii*) *geo-coordinates* of a particular resource of the article.

Similarly, Freebase[3] is a huge online knowledge base which users can edit in a similar manner as Wikipedia. The latest version of Freebase [4] comprises of 85 domains, more than 20 million entities and more than 10 thousand relations across a large number of these domains. In contrast to DBpedia however, in Freebase the source of articles include Wikipedia as well as other sources such as MusicBrainz, WordNet, OurAirports, etc [5]. The classification of articles in Freebase is also slightly different; for a given Freebase article: *i*) a *domain* denote the topic of the article; *ii*) a *type* define a particular kind of entity such as person or location (for e.g. "Lady Gaga" is a Person); and *iii*) *properties* describe an entity (for e.g. "Lady Gaga" has a "place of birth"). Another notable difference between the two knowledge source is the level of deepness in the hierarchy for a particular category or topic.

3 Related Work

DBpedia and Freebase KSs have been important knowledge sources in many classification tasks such as topic detection and semantic linking of Twitter messages. These approaches mostly employ traditional machine learning algorithms building a classifier on Twitter dataset and deriving useful features from KSs.

To date, to the best of our knowledge, no analysis has been done in exploiting these KSs for cross-domain (CD) topic classification of tweets and also in measuring the similarity between these KSs and Twitter. In the following section we thus provide a summary of the related work using these KSs for Twitter on topic detection and semantic linking.

Related Work on using DBpedia for Topic Classification of Tweets Ferragina et al. [7] propose the TAGME system, which enriches a short text with Wikipedia links by pruning n-grams unrelated to the input text. Milne et al. [11] propose an automatic cross-reference of Wikipedia documents and Wikipedia links by means of machine learning classifiers. This method has been shown to not perform well when applied to tweets [10]. Munoz et al [1] also address the problem of assigning labels to microposts, in order to identify what a micropost is about. In their approach they assign DBpedia resources to post by means of a lexicon-based similarity relatedness metric.

Meij et al [10], also assign resources to microposts. In their approach they make use of Wikipedia as a knowledge source, and consider a Wikipedia article as a *concept*,

[3] http://www.freebase.com/

[4] `http://download.freebase.com/datadumps/2012-07-19/`

[5] `http://wiki.freebase.com/wiki/Data_sources`

their task then is to assign relevant Wikipedia article links to a tweet. They propose a machine learning approach which makes use of Wikipedia n-gram and Wikipedia link-based features. Our approach differs from theirs in two main points: 1) rather than considering a Wikipedia article or DBPedia resource link as a concept, we consider a whole DBpedia category as a concept; 2) our study analyses the use of DBpedia as an annotated source dataset, which can be used to increase the performance of machine learning classifiers for assigning a topic label to a tweet.

Mendes et al. [12] propose the Topical Social Sensor, which is a system that allows users to subscribe to hashtags and DBpedia concepts in order to receive updates regarding these topics. They link a tweet with the DBpedia concepts derived from the entities contained in it. This system is designed for detecting a hype on a topic defined a priori. In our work rather than relating a tweet with the DBpedia concepts derived from named entities, we propose the use of DBpedia articles to model a category, and perform an use this articles as source dataset for training a topic classifier to assign a topic label to a tweet.

Related Work on using Freebase for Topic Classification of Tweets Kasiviswanathan et al[9] propose a detection-clustering based approach for streamed topic detection they make use of entities and their types gathered from Freebase. In this paper, rather than proposing a new approach for topic detection we compare the performance of two classifiers; one based on DBpedia and the other on Freebase for detecting topics of tweets.

4 Methodology

This section describes three different steps required for the analysis presented in this paper. The first step, described in Section 4.1, consists on the compilation of datasets from KSs; the second step, described in Section 4.2, consists on the use of these datasets for the development of CD topic classifiers; and the third step consists on the introduction of similarity metrics that can characterise distributional changes between datasets.

4.1 Collecting Data from KS

In this section we refer to our datasets, which will be further described in Section 5. The Twitter dataset consists of a collection of tweets, which were annotated with 17 different topics using the OpenCalais services. In order to compile a set of articles relevant to each of these 17 topics, from both DBpedia and Freebase KSs, we performed two steps. In the case of DBpedia, for a given topic, we SPARQL[6] queried for all resources whose categories and subcategories are similar to the topic. For the returned resources we only kept the first 500 characters from the resources' abstracts. In the case of Freebase, we downloaded the articles using the Freebase Text Service API[7]. Given a topic, we collected all the articles whose domain matched the topic [8]. In addition, for some of the

[6] http://www.w3.org/TR/rdf-sparql-query/

[7] `http://wiki.freebase.com/wiki/Text_Service`

[8] The collection of domains are enumerated at `http://www.freebase.com/schema`

topics (e.g. Disaster or War), which were not defined as domains in Freebase, we looked at all the articles containing these topics in their title. While this service also allows to download the full content of an article, similarly to DBpedia, we only considered the first paragraph up to 500 characters.

The following Subsection 4.2, describes how the DBpedia and Freebase datasets are used to built three different CD classifiers for detecting topics in tweets.

4.2 Building Cross-Domain(CD) Topic Classifier of Tweets

We formally describe each dataset D as a tuple $(X, F, P(X))$ composed of a set of *instances* X, a set of *features* F and a *marginal probability distribution* $P(X)$. Each instance $x \in X$ is represented by a vector of features $x = (f_1, .., f_m), f_i \in F$. The possible *topics* $y = \{cat_{Y_1}, \ldots, cat_{Y_d}\}$ for an instance x can take values from $Y \in \{cat_1, \ldots, cat_k\}$. The goal of the classification then is to learn a model $h : X \rightarrow Y$ from a set of annotated *training data* $L = \{(x_1, y_1), \ldots, (x_n, y_n) | x_i \in X, y_i \in Y\}$, that induces a non-injective map between X and Y such that multiple class labels can be assigned to the same instance - e.g. $h(x_1) = \{cat_1, cat_2\}$.

A CD scenario consists of a *source dataset* $D_S = (F_S, X_S, P(X_S))$ –on which the classifier is built– and a *test dataset* $D_T = (F_T, X_T, P(X_T))$ –on which the classifier is evaluated–. As illustrated in Figure 1, in this paper we consider three cases for the source dataset. The first two cases aim to investigate the usefulness of DBpedia and Freebase KSs independently, and the third case combines the contribution of both KSs. Thus, the CD scenarios studied in the paper are described as follows: **Scenario I (Sc.DB)** consisting of *sole DBpedia articles*; **Scenario II (Sc.FB)** consisting of *sole Freebase articles*; and **Scenario III (Sc.DB-FB)** consisting of a *joint set of DBpedia and Freebase articles*. The test dataset in each case is the Twitter dataset.

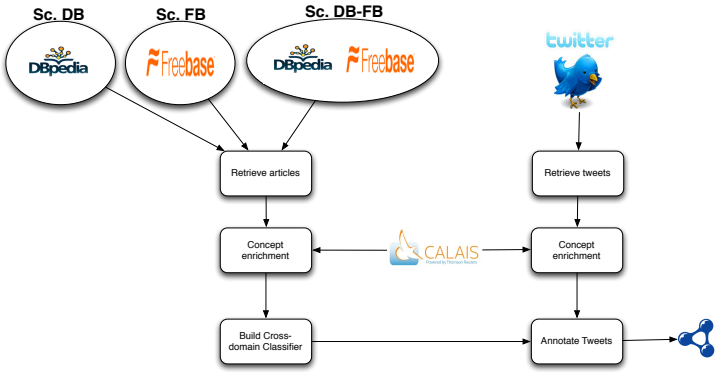

Fig. 1. The Sc.DB, Sc.FB and Sc.DB-FB CD scenarios using concept enrichment.

We used as baseline classifier an SVM classifier with linear kernel, which has been found to perform best for transfer learning [6]. We also took the commonly used one-

vs-all approach to decompose our multi-label problem into multiple independent binary classification problems.

Feature Extraction The performance of the machine learning algorithm rely on the feature representation employed. We propose two different feature sets for the examples in both train and test datasets:

- *a bag-of-word*(***BoW***) representation: This representation captures our natural intuition to utilise what we know about a particular topic, so that the features which are most indicative of a topic can be detected and the appropriate label(s) assigned. This feature consists of a collection of words weighted by TF-IDF (term frequency-inverse document frequency) in order to capture the relative importance of each word.
- *a bag-of-entities*(***BoE***) feature representation. The second set of features makes use of named entities. These entities were extracted by querying OpenCalais API[9] for entity extraction on each instance belonging to the Dbpedia, Freebase and Twitter datasets as presented in Figure 1. We then used the dereferenceable URI and concepts returned by the API as features for the classifier. Our intuition is that entities can be characteristic of a topic, serving as trigger words for this topic; reducing in this way the lexical differences between the source and target datasets.

4.3 Measuring Distributional Changes Between KS and Twitter

In addition to building the CD classifiers, we investigated various measures for quantifying the similarity between KSs and Twitter. When building a machine learning classifier, it is expected that the closer the train dataset to the test dataset the better the performance of the classifier [13]. Therefore, these similarity metrics can be potentially useful in predicting the adequacy of the data collected from a KS in detecting topics in tweets.

To measure the similarity between the distributions of the presented datasets, let $\overrightarrow{d}$ represent a vector consisting of all the features occurring on a dataset. Then, $\overrightarrow{d_s}$ denotes such a vector for the train dataset and $\overrightarrow{d_t}$ for the test dataset. In light with the feature set employed, the $\overrightarrow{d_s}$ and $\overrightarrow{d_t}$ contain the TF-IDF weight for either the **BoW** or **BoE** feature sets. Then the proposed statistical measures are:

- the *chi-squared (χ^2) test*: The χ^2 test measures the independence between the feature sets (F_S and F_T) and the train and test datasets. Given the $\overrightarrow{d_s}$ and $\overrightarrow{d_t}$ vectors , the χ^2 test can be computed as

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

, where O is the observed value for a feature, while E is the expected value calculated on the basis of the joint corpus.

[9] www.opencalais.com/

– the *Kullback-Leibler symmetric distance* (KL): Originally introduced in [3], the symmetric KL divergence metric measures how different the $\overrightarrow{d_s}$ and $\overrightarrow{d_t}$ vectors are on the joint set of features $F_S \cup F_T$:

$$KL(\overrightarrow{d_s}||\overrightarrow{d_t}) = \sum_{f \in F_S \cup F_T} (\overrightarrow{d_s}(f) - \overrightarrow{d_t}(f)) \log \frac{\overrightarrow{d_s}(f)}{\overrightarrow{d_t}(f)}$$

– *cosine similarity* measure: The cosine similarity represents the angle that separates the train and test vectors $\overrightarrow{d_s}$ and $\overrightarrow{d_t}$:

$$cosine(\overrightarrow{d_s}, \overrightarrow{d_t}) = \frac{\sum_{k=1}^{F_S \cup F_T} (\overrightarrow{d_s}(f_{S_k}) \times \overrightarrow{d_s}(f_{T_k}))}{\sum_{k=1}^{F_S \cup F_T} (\overrightarrow{d_s}(f_{S_k}))^2 \times \sum_{k=1}^{F_S \cup F_T} (\overrightarrow{d_t}(f_{T_k}))^2}$$

We also note that some of these proposed functions measure actual similarity ($cosine$), while others measure distance KL, χ^2.

5 Dataset and Data Pre-Processing

The Twitter dataset consists of tweets posted between October 2010 and January 2011, and was originally collected by [2],[10] comprising more than 2 million Tweets posted by more than 1619 users. We further annotated this data set with topics returned by the OpenCalais service, which label each tweet with one or more topics (from a collection of 17 topics). For our analysis we randomly selected one thousand tweets for each topic, excluding re-tweets, resulting in a collection of 12,412 Tweets. Some of these categories are presented in Table 1. Similarly from DBpedia and Freebase we randomly selected one thousand articles for each topic, comprising of 9,465 articles from DBpedia and 16,915 articles from Freebase.

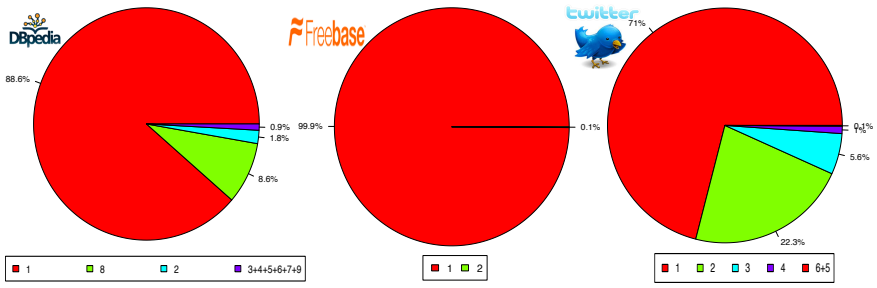

Fig. 2. The multi-label distribution in DBpedia, Freebase and Twitter datasets. The numbers in the legend indicate the number of topics assigned to an example, varying from 1 to 9 topics.

[10] Available at http://wis.ewi.tudelft.nl/umap2011/

In line with previous approaches ([2]), for the datasets we removed all the stopwords and we converted all words into lower case; after which a Lovins stemmer was applied. In addition, in order to reduce the vocabulary differences between the KS datasets and Twitter, all hashtags, mentions and URL links, which are particular to the Twitter dataset, were removed. The feature space was also reduced to the top-1000 words weighted by TF-IDF for each category.

Figure 2 shows the distribution of the examples belonging to multiple topics in each dataset. The Twitter dataset contain some tweets annotated with up to six categories, with the majority of them being annotated with only one topic. In the case of the Freebase dataset, due to the nearly flat hierarchical structure of the domains, the majority of the articles belong to a single category. In the case of the DBpedia dataset the majority of the articles belong to a single category, and less than 1% of the articles are annotated with 3,4,5,6,7 or 9 topics. The size of the vocabulary for each category and

Topic name	*Example tweets*
Business&Finance(*BusFi*)	visa cyber attack transactions affected account data safe company nbc
Disaster&Accident(*DisAcc*)	happening accident people dying could phone ambulance wakakkaka xd
Education(*Edu*)	read book even final pass pages period read
Environment(*Env*)	good complain cause part energized midterms happening
Entertainment&Culture(*EntCult*)	google adwords commercial greeeat enjoyed watching greeeeeat day
Health&Medical&Pharma(*Health*)	unprocessed fat eat lose fat real butter coconut oil eggs olive oil avocados raw nuts
Politics(*Pol*)	quoting military source sk media reports deployed rocket launchers decoys real
Sports(*Sports*)	ravens good position games left browns bengals playoffs
Technology&Internet(*TechIT*)	iphone cute ringtone download ringtone; lets enjoy wikileaks tomorrow publish direct message ever
War&Conflict(*War*)	nkorea prepared nuclear weapons holy war south official tells state media usa

Table 1. Example tweets for some of the evaluated topics after preprocessing (removing the stopwords, hastags, mentions and URLs).

dataset is presented in Figure 3. This distribution presents a variation in the vocabulary size between the different datasets. Namely, in the DBpedia dataset each category is featured by a large number of words. This is expected, as the DBpedia articles are typically longer than the Freebase articles. The richest topics in DBpedia being *Religion*, *EntCult*, *TechIT*. In contrast, in the Freebase dataset the topics are being described by less words. The richest topics in Freebase are *Sports*, *TechIT*, *HumInt*. While for the Twitter dataset these topics are *Env*, *DisAcc*, *BusFi*.

When looking at the frequency of the entities in Figure 4, we can observe similar trends. The DBpedia articles contain the most number of entities for each topic, on average 22.24 ± 1.44 entities. From the full collection 69(0.72%) of the articles do not have any entity. In the case of Freebase, the average number of entities per article is

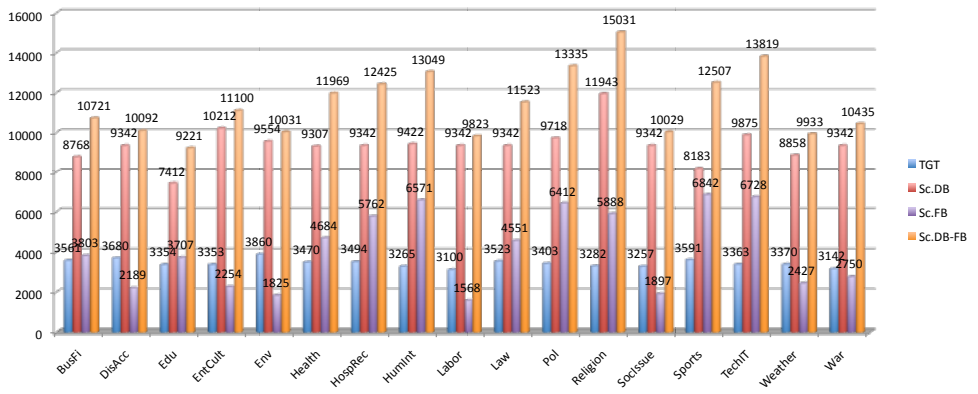

Fig. 3. The size of vocabulary in the source (Sc.DB, Sc.FB, Sc.DB-FB) and target (TGT) datasets after pre-processing.

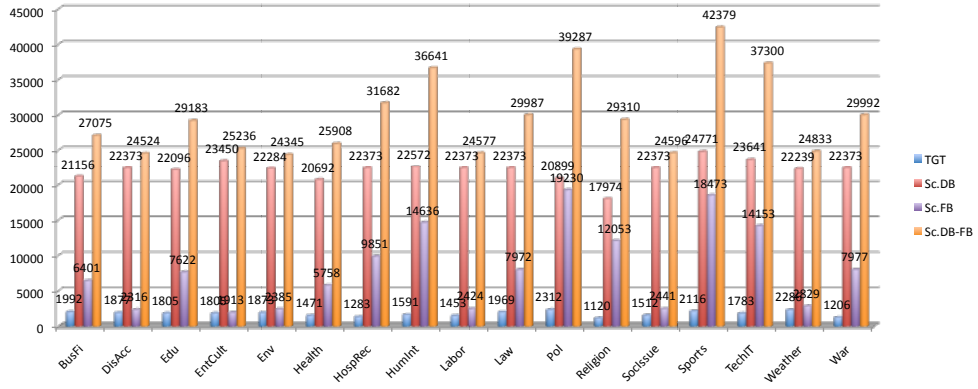

Fig. 4. The number of entities in the source (Sc. DB, Sc. FB, Sc. DB-FB) and target (TGT) datasets after pre-processing.

8.14 ± 5.78. The percentage of articles without any entity is 19.96%(3,377 examples). Lastly, the Twitter dataset contains the smallest number of entities, on average 1.73 ± 0.35 entities per articles. The number of articles mentioning no entity is 5,137 (41%).

The heatmap in Figure 5 demonstrates how the entity types' frequencies differ across different datasets. The darker the color, the higher the frequency of an entity in a dataset. According to this figure, Organization and Position have a relatively high frequency across all datasets. Other entities appearing frequently on these datasets are Person, Country and Natural Feature. Entity types such as MedicalCondition, or SportsEvent appear to be more representative of particular topics such as *Health* and *Sports*. When checking the clustering by topic in Figure 5, we can notice that the *Health* and *Edu* topics present a similar entity distribution in DBpedia and Freebase; the *War* topic

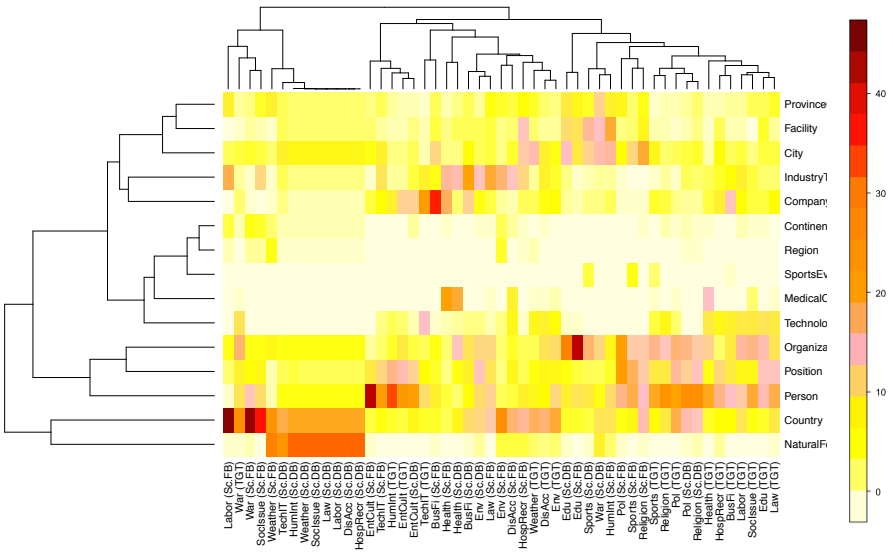

Fig. 5. The distribution of the top 15 entity types the in the source (Sc. DB, Sc. FB) and target (TGT) datasets.

has a similar entity distribution in Twitter and Freebase; while the *Pol* category presents a similar entity distribution in Twitter and DBpedia.

Based on the above figures on lexical richness and entity frequency, thus, we can notice that the Freebase dataset exhibits more similarity to Twitter than Dbpedia datasets. In order to get a better insight into this similarity, we will compare these datasets according to the proposed measures in the coming section.

6 Experiments

In this section we perform a series of experiments to investigate which KS exhibits more similarity to Twitter. In the first set of experiments we compare the performance of the SVM classifiers derived for the proposed cross-domain (CD) scenarios (Subsection 4.2), with the SVM classifier built on Twitter data only. These classifiers were trained using the different BoW and BoE features(Section 6.1) in each scenario. Therefore in this first set of experiments we address the questions of *which KS reflects better the lexical variation in Twitter?* and *what feature makes the KSs look more similar to Twitter?*.

The second set of experiments, consists on computing the correlation between the proposed statistical measures (Section 6.2) and the accuracy of the CD classifiers. In this correlation analysis we investigate *which statistical measure presents the highest correlations with the accuracy of a CD classifier?* providing the most reliable estimate for the quality of KSs in topic classification of tweets.

6.1 Comparison of the Different Feature Sets in Cross-Domain Scenarios

The SVM classifiers derived from the CD scenarios –**Sc.DB**, **Sc.FB** and **Sc.DB-FB**– were evaluated based on their performance when trained using **BoW** and **BoE** features. The **TGT** SVM classifier –based on Twitter data only– was built on 80% of the Twitter data, and evaluated on 20% of the twitter data over five independent runs.

Figure 6 shows the results obtained using **BoW** and **BoE** features for the different CD scenarios. Based on the average performance in F1 measure, we can observe, that among the three CD scenarios, the best average performance was obtained by the **Sc.DB-FB** SVM classifier using **BoW** features, which is followed by the **Sc.FB** and **Sc.DB** SVM classifiers also using **BoW** features.

Using both feature sets, we found that for the **Sc.DB-FB** scenario the topics which presented a performance closer to the one obtained by the **TGT** classifier were the *Weather* and *Edu*. For the **Sc.FB** scenario these topics were the *Edu*, *Weather*, *Labor*. Finally for the **Sc.DB** scenario these topics were the *Edu*, *Health*. The topics for which the performance was higher using **BoE** features were the *BusFi*, *Env*, *Pol*, *SocIssue*, *Sports*. For *Labor* the performance was the same for both features .

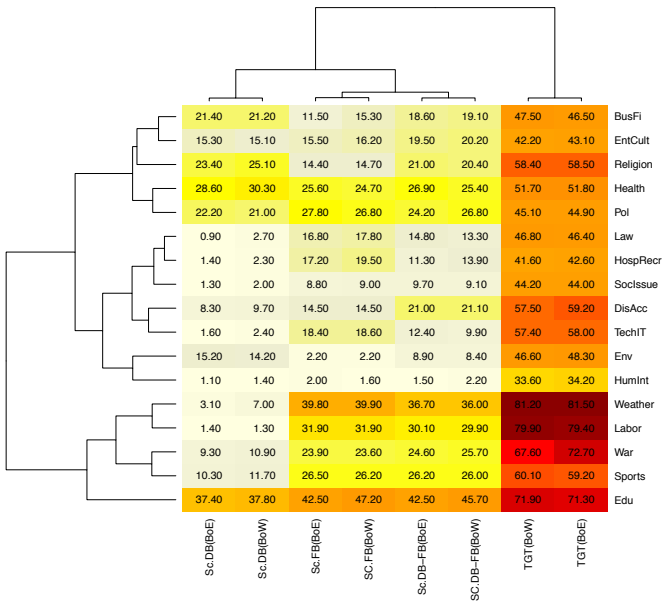

Fig. 6. The performance in F1 measure of the Sc.DB, Sc.FB, Sc.DB-FB and TGT classifiers using **BoW** and **BoE** features for each topic over five independent runs. The training set of TGT classifier consists of 80% of the Twitter dataset (9,928 tweets). The Sc.DB, Sc.FB and Sc.DB-FB classifier were trained only on social knowledge sources data.

A slightly different trend can be observe for the **TGT** classifier, where the best average F1 measure was achieved using **BoE** features. There were 10 topics for which

BoE features were useful: *DisAcc*, *EntCult* , *Env*, *Health*, *HospRec*, *HumInt*, *Religion*, *TechIT*, *War* and *Weather*.

Overall, our results indicate that **Sc.FB** KS is more similar to Twitter than **Sc.DB**. Furthermore, combining the contribution of the **Sc.DB** and **Sc.FB** is beneficial for detecting topics in Tweets, since the **Sc.DB-FB** scenario achieves the best overall results. With regard to the features, we found that in 11 out of 17 cases the results obtained using **BoW** features were better, and in 5 out of 17 cases the **BoE** features were found more effective.

We also compared the performance of the Twitter classifier against the three CD classifiers over the full learning curve, by gradually increasing the number of tweets used to train the classifier. Our analysis revealed that in the majority of the cases the CD classifiers worked relatively well. That is, a sufficient amount of annotated tweets were needed to significantly outperform the three CD classifiers over the full learning curve. The number of annotations needed for each topic is summarised in Table 2. For e.g. for more than 9 out of 17 topics the necessary amount of annotated tweets need to exceed 900. However, in a real-world scenario annotating tweets is an expensive task.

BusFi	*DisAcc*	*Edu*	*EntCult*	*Env*	*Health*	*HospRec*	*HumInt*	*Labor*
$993^{\clubsuit}$	$993^{\clubsuit}$	$993^{\clubsuit}$	$993^{\clubsuit}$	$993^{\clubsuit}$	$1,986^{\clubsuit}$	$1,986^{\spadesuit}$	$160^{\spadesuit}$	$320^{\spadesuit}$
Law	*Pol*	*Religion*	*SocIssue*	*Sports*	*TechIT*	*Weather*	*War*	
$640^{\spadesuit}$	$993^{\clubsuit}$	$320^{\spadesuit}$	$320^{\spadesuit}$	$993^{\spadesuit}$	$640^{\clubsuit}$	$320^{\spadesuit}$	$640^{\clubsuit}$	

Table 2. Number of annotated tweets required for the Twitter classifier to beat the Sc.DB, Sc.FB and Sc.DB-FB CD classifiers. Significance levels: p-value $< \clubsuit 0.01 \spadesuit 0.05$

6.2 Comparison of Statistical Measures in Topic Classification of Tweets

In this second set of experiments we aimed to investigate our research question of *how similar or dissimilar are social knowledge sources to Twitter posts; and which similarity measure does better reflect the lexical changes between KSs and Twitter posts?*. We thus performed a comparison between the proposed KL divergence, $cosine$ similarity and χ^2 test by measuring the correlation of these values with the performance of a CD classifier using Sc.DB, Sc.FB and Sc.DB-FB scenarios.

Each CD classifier was evaluated on 20% of the Twitter data, and the performance was averaged over five independent runs. The obtained F1 measures for the CD classifiers were then compared with the values obtained for the different statistical measures, and the Pearson correlation was computed.

Figure 7 show the correlations obtained using KL, cosine and χ^2 values. A positive correlation indicates that the performance of the CD classifiers increases as the divergence decreases (the distribution are more similar); while a negative correlation indicates that the performance increases as the divergence increases (the distributions are less similar). As we can notice, for the KL scores, there are 24 cases in which the correlation scores are higher than 70% in absolute terms. In the case of $Cosine$ similarity these cases sum up to 25. While in the case of χ^2 values for a total of 32 cases were the correlation values higher than 70%.

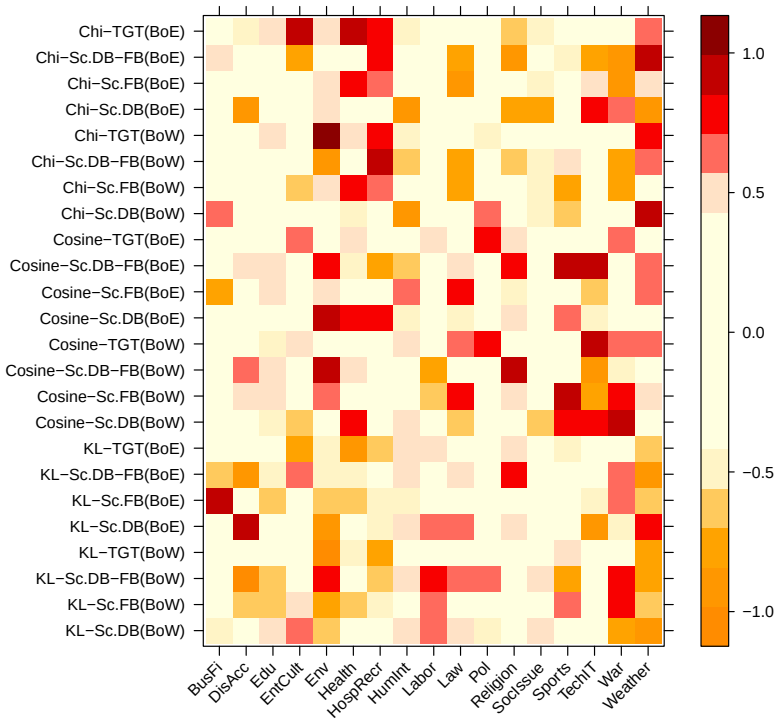

Fig. 7. The Pearson correlation between the performance in F1 of the Sc.DB, Sc.FB, Sc.DB-FB CD classifiers and the KL, Cosine and χ^2 measures

Based on these results, we found the χ^2 to provide the best correlation scores for the usefulness of the KSs data. The second best score was for the $cosine$ similarity, which was followed by the KL measure.

Figure 8 shows the pairwise similarity obtained for the source and target datasets according to $(\chi^2)^{-1}$ similarity measure.[11] As expected the closest datasets to the test Twitter dataset is the training set for the Twitter classifier (ChiSc.TGT). The second closest dataset according to χ^2 is the Sc.Fb dataset. The Sc.DB and Sc.DB-FB are then the less similar datasets to the test dataset.

7 Conclusions and Future Work

In this paper we presented a first attempt towards understanding the usefulness of DBpedia and Freebase KSs in CD topic classification of tweets. We presented an analysis

[11] As χ^2 measure distance rather than similarity we inverted its value to present the similarity between topics better.

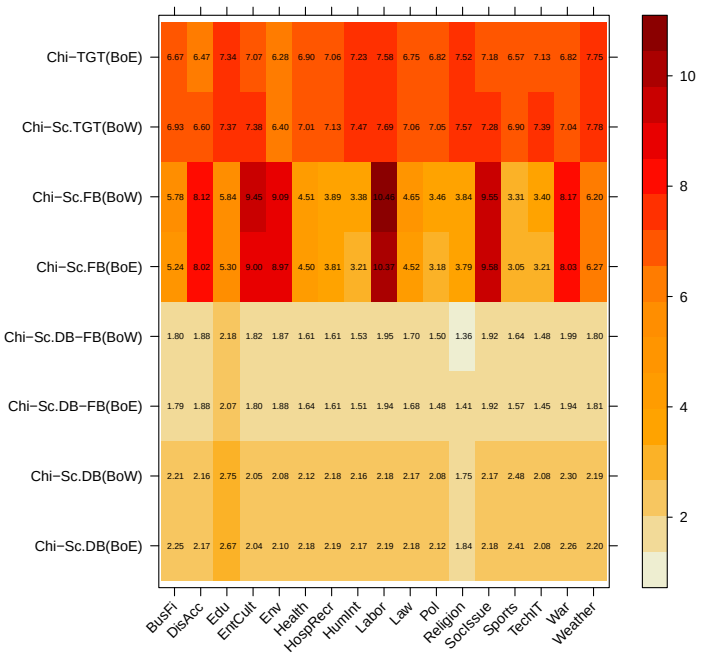

Fig. 8. The values of $(\chi^2)^{-1} * 10^{-5}$ for each Sc.DB, Sc.FB, Sc.DB-FB, TGT scenarios. High values indicate that the topics are more similar between the source and target dataset.

between these data sources focusing on various lexical features (**BoW**) and entity features(**BoE**).

For a total of 17 topics we compiled a gold standard for each individual KS, and for the joint set of these sources. From the resulted datasets we then built three CD classifiers which we evaluated against a Twitter classifier using the different features.

Our analysis revealed that from the two KSs, Freebase topics seem to be much closer to the Twitter topics than the DBpedia topics due to the much restricted vocabulary used in Freebase. Furthermore, we found that the two KSs contain complementary information, i.e.; the joint dataset was found more useful than the individual KS datasets. With regard to the feature sets, we found that for the three CD classifiers on average the results obtained using **BoW** were better than those obtained with **BoE** in 5 out of 17 cases.

When comparing the results of these CD classifiers to the Twitter classifier we found that for some of the topics the Twitter classifier required a large number of annotations to outperform these classifiers, indicating that in the absent of any annotated tweets, applying these CD classifiers is still beneficial. Previous research on transfer learning has also shown, that outperforming the target (Twitter) classifier is extremely difficult for many tasks including sentiment classification ([5, 13]). A promising alternative

found in the literature was to combine the annotated examples in the source and target datasets([6]). Our future work aims to follow this direction, focusing on building transfer learning algorithms which can effectively combine the contribution of the two KSs; and also exploring other features derived from the named entities.

Finally, we also looked at various statistical measures for predicting the usefulness of the data gathered from these KSs. These experiments revealed the χ^2 test as being the best measure for quantifying the distributional differences among between KSs and Twitter. Our future work in this direction will focus in investigating more accurate measures for quantifying this difference for e.g. by taking into account the special vocabulary (abbreviations, misspellings, shortening) used in Twitter, and normalise this to standard English terms ([8]).

References

1. Identifying topics in social media posts using dbpedia. In O. e. a. Munoz-Garcia, editor, *In Proceedings of the NEM Summit (27-29 September 2011)*, pages 81–86, 2011.
2. F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on twitter for personalized news recommendations. In *Proceedings of the 19th international conference on User modeling, adaption, and personalization*, UMAP'11, pages 1–12, Berlin, Heidelberg, 2011. Springer-Verlag.
3. B. Bigi. Using kullback-leibler distance for text categorization. In *Advances in Information Retrieval, Lecture Notes in Computer Science Volume 2633*, 2003.
4. C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. Dbpedia - a crystallization point for the web of data. *J. Web Sem.*, 7(3):154–165, 2009.
5. J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Association for Computational Linguistics*, Prague, Czech Republic, 2007.
6. H. Daume III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.
7. P. Ferragina and U. Scaiella. Tagme: on-the-y annotation of short text fragments (by wikipedia entities). In *Proc of the CIKM'10*, 2010.
8. B. Han and T. Baldwin. Lexical normalisation of short text messages: makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 368–378, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
9. S. P. Kasiviswanathan, P. Melville, A. Banerjee, and V. Sindhwani. Emerging topic detection using dictionary learning. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 745–754, New York, NY, USA, 2011. ACM.
10. E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining, WSDM '12*.
11. D. Milne and I. H. Witten., editors. *Learning to link with Wikipedia.* 2008.
12. P. K. P. N. Mendes, A. Passant and A. P. Sheth. Linked open social signals. In *In WI-IAT 10*, 2010.
13. S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010.